Data-Centric Optical Networks and Their Survivability

Didier Colle, Sophie De Maesschalck, Chris Develder, Pim Van Heuven, Adelbert Groebbens, Jan Cheyns, Ilse Lievens, Mario Pickavet, Paul Lagasse, and Piet Demeester

***Abstract*—The explosive growth of data traffic—for example, due to the popularity of the Internet—poses important emerging network requirements on today's telecommunication networks. This paper describes how core networks will evolve to optical transport networks (OTNs), which are optimized for the transport of data traffic, resulting in an IP-directly-over-OTN paradigm.**

Special attention is paid to the survivability of such data-centric optical networks. This becomes increasingly crucial since more and more traffic is multiplexed onto a single fiber (e.g., 160×10 Gb/s), implying that a single cable cut can affect incredible large traffic volumes. In particular, this paper is tackling multilayer survivability problems, since a data-centric optical network consists of at least an IP and optical layer. In practice, this means that the questions "In which layer or layers should survivability be provided?" and "If multiple layers are chosen for this purpose, then how should this functionality in these layers be coordinated?" have to be answered.

In addition to a theoretical study, some case studies are presented in order to illustrate the relevance of the described issues and to help in strategic planning decisions. Two case studies are studying the problem from a capacity viewpoint. Another case study presents simulations from a timing/throughput performance viewpoint.

***Index Terms*—Multilayer survivability, MPλS, multiprotocol label switching (MPLS), IP-over-OTN, recovery, capacity dimensioning.**

I. Introduction

THIS paper begins by explaining in Section I-A how current networks are rapidly evolving toward data-centric optical networks. Section II continues with an overview of proposed recovery techniques for multiprotocol label switching (MPLS) networks and how these techniques can be adopted in the optical layer in case of a generalized (G-)MPLS network. A novel technique called *fast topology-driven constraint-based rerouting* is also proposed in that section. The core of this paper deals with the issue of providing survivability in a multilayer network. Several approaches are discussed theoretically in Section III. This section concludes with a first indication for the particular case of an IP-MPLS directly over optical transport network (OTN)-MPλS network. Section IV continues by means of some case studies concentrating on the investigation of specific issues (i.e., capacity dimensioning and the impact of fast protection switching on transmission control protocol (TCP) behavior). Final conclusions are presented in Section V.

Manuscript received March 25, 2001; revised July 30, 2001. This work was supported in part by the European Commission under the IST-Projects LION, TEQUILA, and DAVID and in part by the Flemish Government under IWT-Project ITA/980272/INTEC. The work of P. Van Heaven was supported by the Flemish government under an IWT-scholarship. The work of C. Develder, A. Groebbens, and J. Cheyns was supported by the Fund for Scientific Research—Flanders, Belgium.

The authors are with the Department of Information Technology, Ghent University—IMEC, 9000 Gent, Belgium (e-mail: didier.colle@intec.rug.ac.be; sophie.demaesschalck@intec.rug.ac.be; chris.develder@intec.rug.ac.be; pim.vanheuven@intec.rug.ac.be; adelbert.groebbens@intec.rug.ac.be; jan.cheyns@intec.rug.ac.be; ilse.lievens@intec.rug.ac.be; mario.pickavet@intec.rug.ac.be; lagasse@intec.rug.ac.be; demeester@intec.rug.ac.be).

Publisher Item Identifier S 0733-8716(02)00164-6.

A. From IP/ATM/SDH/WDM to IP-MPLS Directly Over OTN-MPλS

The popularity of the Internet [1], [2] has lead in recent years to an explosive growth in the traffic to be carried by telecommunication networks. Data traffic even dominates voice traffic [3], and recent forecasts do not seem to predict a quick slowdown [3], [4].

It is obvious that this will have a major impact on today's telecommunication networks. These networks will be increasingly optimized for the dominant data (mainly IP) traffic. Today, a typical (core of a) telecommunication network consists of a transport network carrying the traffic of several parallel services: e.g., plain-old switched telephone service, leased-line services, etc. Such a transport network (TN) may, for example, consist of an asynchronous transfer mode (ATM) network (functioning as service integration layer) on top of a synchronous digital hierarchy (SDH) network. Fiber exhaust is currently solved by multiplying the capacity of a fiber ten—or even a hundred—times by means of point-to-point wavelength-division multiplexing (WDM) systems. Recently, WDM systems of 160 10-Gb/s wavelengths have been announced [5]. This multiplexing technique has proven to be very cost-efficient due to the economy-of-scale [6].

It is obvious that incumbent operators also want to profit from the new Internet service provider (ISP) market fragment. They are at a more comfortable position, since they still have their important revenue-generating voice [3] business and other services, in contrast to newcomers. However, they are of course not willing to immediately replace their current infrastructure, and thus they start their ISP business by running their IP network in parallel with their currently existing network services, on top of the same transport network. This means they typically are in (or have just left) an IP/ATM/SDH/WDM multilayer scenario [7]. The practical meaning of this scenario is explained in Fig. 1.

The transport of IP packets through ATM has some major drawbacks. First, there is the important cell tax: approximately 10% overhead (5 bytes header per 48 bytes payload). Secondly, an IP packet has a typical length of 500 or 1500 bytes [8] and is thus typically encapsulated in many ATM cells. This implies

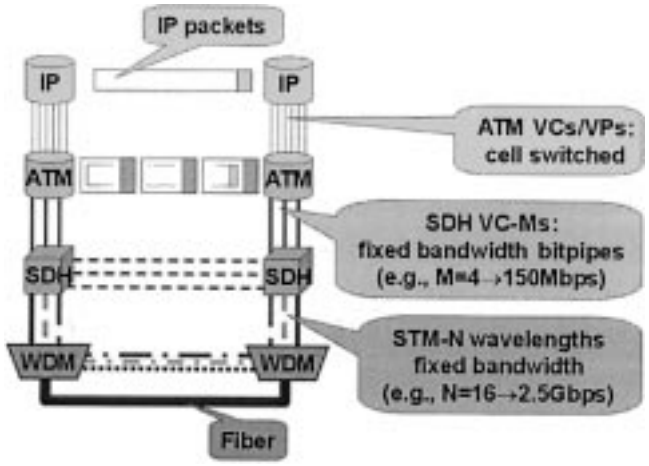

Fig. 1. Illustration of the IP/ATM/SDH/WDM technology mapping. IP routers exchange IP packets by sending them through ATM connections, which requires encapsulation of an IP packet in many ATM cells. ATM nodes are interconnected by fixed-bandwidth bitpipes (VC-Ms) through the SDH network. The capacity on the fibers interconnecting the SDH DXCs is increased by multiplexing multiple wavelengths onto a single fiber.

that per IP packet, many ATM cells have to be handled and processed in intermediate ATM nodes. Yet another disadvantage is that there is an extra layer to maintain and manage. Of course, ATM also has its benefits: its connection orientation, opening opportunities for traffic engineering (TE), due to the decoupling of routing (control plane) and forwarding (data plane).

However, the steady and ongoing progress and research in optimizing IP router designs [9] implies that IP does not have to take the drawbacks of ATM for granted, if it was able to overcome its lack in TE-capability. The MPLS concept, grown within the IETF, has proven to be suitable for this purpose [10]–[13]. Thus, in the end, we may expect that an MPLS-empowered IP network absorbs the TE-feature of ATM and bypasses the ATM layer by coding the MPLS labels in a shim-header in front of the IP packet. Similar to ATM, a label switched router (LSR) will label-switch the packets (i.e., look up the incoming ⟨interface, label⟩-pair in the label information base (LIB), in order to know along which interface to forward the packet with which label). This bypasses the legacy and cumbersome lookup operations of the destination address in the routing table. To populate the LIB with appropriate mapping information, a protocol [either the label distribution protocol (LDP [14]) or the resource reservation protocol (RSVP [15])] in the MPLS control plane will be used, allowing one to set up and tear down so-called label switched paths (LSPs) through the MPLS network. (Note that in the remainder of this paper, we will use the following terminology: *IP network* refers to an MPLS-incapable network, *MPLS network* is short for an MPLS-capable IP network, and *IP-MPLS network* will be used when it can be either an IP or an MPLS network. It also may happen that we call an MPLS network an MPLS-empowered/capable IP network (to stress the MPLS capability). The services and traffic (demand) carried by an IP-MPLS network are always indicated by *IP services* and *IP traffic*, respectively.)

Even more, the steady growth of the IP traffic will (soon) allow bypassing the ATM-layer, simply because the SDH switching granularity will match the required line-speeds for the direct interconnection of IP-MPLS routers. IP-MPLS-router interface cards of up to 622 Mbps or even 2.5 Gb/s are currently commercially available and deployed [9], [16], [17]. As traffic will not stop growing, in no time SDH digital cross-connects (DXCs) will not be able to catch up with the required switching granularity (a coarse granularity of the underlying layer is beneficial for the IP-MPLS network from a scalability point of view). At that moment, the SDH network will be bypassed as well: SDH switching and multiplexing will not be used anymore (but SDH framing may still be used). Instead, the cross-connect functionality will be pushed into the optical domain, resulting in an OTN. Optical network elements (ONEs) with limited flexibility are already commercially available, and fully flexible large optical cross connects (OXCs) are ready for massive commercialization [5], [18].

A final consideration in our roadmap for next-generation networks is the fact that transport networks tend to be rather static, due to the fact that an operator has to set up each connection manually through the network management system (NMS). This does not match with the exponentially growing and highly dynamic IP traffic pattern, requiring frequent changes of the wavelength bandwidth pipes provisioned by the OTN network to carry the IP-MPLS network traffic. Therefore, a current hot research topic is to investigate how this provisioning process can be automated. As in all switched networks, the control plane will serve this need, as illustrated in Fig. 2. Signaling through the control channel of the user–network interface (UNI)—and thus between the IP-MPLS and OTN network—(e.g., OIF UNI spec 1.0 [19]) makes it possible for the client to automatically request the setup of a new lightpath through the OTN. The control channel through the network–network interface (NNI) allows the exchange of signaling messages for routing protocol information exchange [e.g., link-state advertisements (LSAs) being used in the open shortest path first (OSPF) routing protocol], setup of a lightpath, etc.

Generally speaking, two main (extreme) models exist for an automatic switched optical network (ASON). ITU-T G.astn [20] targets an overlay model for an automatic switched (optical) transport network (ASTN is a generalization of ASON). In the overlay model, both the transport and its client networks have a separated and independent control plane. The IETF targets more a peer model with the generalized-MPLS (G-MPLS) concept. This concept originated from MPλS, where the idea was that a wavelength (lambda) is a label like any other label and therefore the MPLS concept can be adopted in the optical domain to serve the need for fast automatic lightpath (or optical LSP) provisioning [7], [21]. G-MPLS is generic in the sense that it considers any type of label: a header-bitstring for a packet-switch capable LSR (PSC-LSRs), a time slot for a TDM-switch capable ISR (TSC-LSR: e.g., SDH-DXC), a wavelength for a lambda-switch capable LSR (LSC-LSR: e.g., OXC), or even a fiber in a fiber-switch capable LSR (FSC-LSR) [22]. A similar terminology as the one for IP-MPLS networks will be used for optical networks: OTN refers to an optical network not controlled by MPλS, an MPλS network to an optical network controlled by an MPλS control plane, and an OTN-MPλS network to an optical network, regardless of the type of the control plane.

Although both client and transport networks may have their own separate and independent (G-)MPLS control planes, an in-

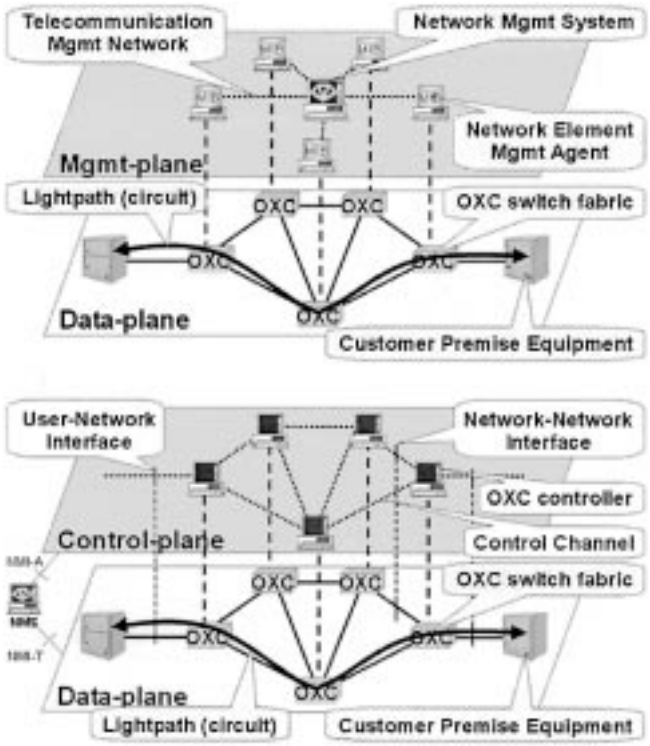

Fig. 2. The difference between a static OTN at the top and an automatic switched optical transport network (ASON) at the bottom of the figure. An ASON is an OTN, empowered with a (distributed) control plane (taking over a large part of the crucial functionality of the management plane), allowing signaling with the client through the UNI, in order to realize a switched optical channel service.

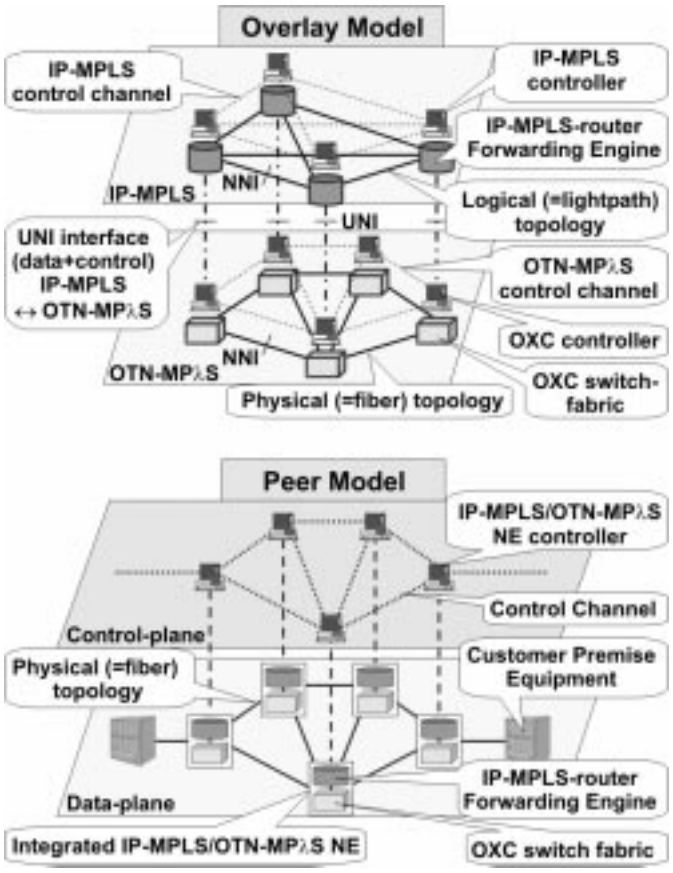

Fig. 3. Illustration showing that in the overlay model (top) the client network is controlled by a separate control plane, independent from the control plane of the transport network. This is in contrast to the peer model (bottom) where the control plane of the client network is integrated into the control plane of the transport network: thus, collocated client and transport network equipment is seen as a single entity.

tegration of those control planes into a single one (covering both layers) seems obvious, resulting in the so-called peer model. The difference between overlay and peer model is illustrated in Fig. 3. The peer model may have some advantages: avoiding duplication of control plane functionality in distinct layers and avoiding the requirement of standardization of a UNI between IP-MPLS routers and OXCs (since the single integrated control plane controls both layers). However, it suffers from the fact that integration and compatibility among multiple client (type) networks seem to be hard and that all information (including confidential information like the TN-topology) is freely accessible in the client domain.

Considering the expectation that in the long term the peer model will become mature enough and eventually overtake the overlay model (when IP-MPLS becomes the service integration layer), we propose as a horizon for our roadmap a peer-modeled IP-MPLS/OTN-MPλS network. Note that this is the horizon of our roadmap, not the end of network evolution. There are already ideas to drive the switching granularity even higher (waveband switching or even fiber switching), and intensive research is going on in the field of optical packet switching.

II. Enhancing Survivability Features of the G-MPLS Technology for IP and OTN Networks

It was already mentioned that the decoupling in MPLS of routing and forwarding opens opportunities for traffic engineering. This is true in particular for the resilience aspects in TE. The goal of this section is to give a brief summary of the current proposals for network recovery in MPLS networks. The impact of G-MPLS is also studied. The reader is referred to [7], [13], and [23]–[31], for more detailed information (terminology is not fixed yet and therefore we use our own terminology in this paper). Note that this section is focusing on resilience in a single layer (thus MPLS or MPλS): multilayer issues are presented in a later section. The section is divided into protection and restoration, referring to whether an alternative path is preestablished or not.

A. Restoration in MPLS

Restoration typically means that connections affected by a failure are routed along an alternative path that is calculated and set up at the time of the failure: a big advantage of restoration is its flexibility. Restoration also allows sharing spare capacity between several failure scenarios. MPLS rerouting is an example of restoration. *MPLS rerouting* relies on the dynamic IP routing protocols. Failures are detected by adjacent routers (e.g., endpoints of a failing link) and advertised/flooded over the network in order to allow other routers to take this topology change into account. After updating its routing tables, a router somewhere in the network may notice that it has LSPs leaving along another interface than indicated by the routing table entries corresponding to the destination of these LSPs. This will trigger

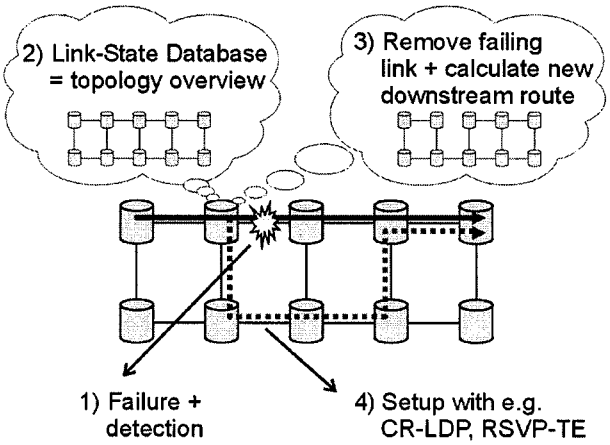

Fig. 4. An LSR detecting a failure will reroute in FTCR outgoing LSPs that are affected by that failure. The LSR can compute an alternative route from itself toward the egress LSR based on its own link-state database and set up the LSP by means of explicit routing (e.g., CR-LDP, RSVP-TE) in order to overcome the problem that other LSRs may not be aware of the failure yet.

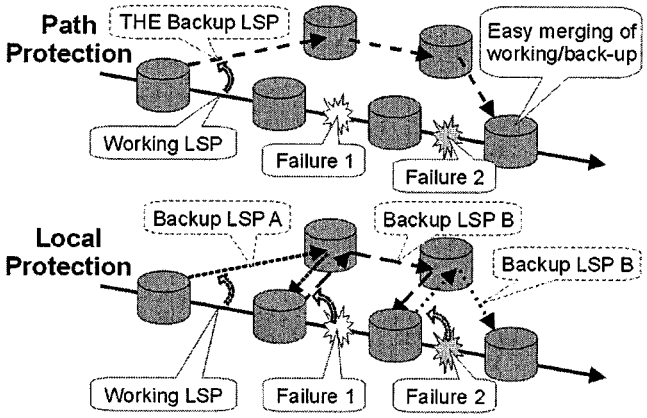

Fig. 5. Explanation of path (top) and local (bottom) MPLS protection under two different failure scenarios. Path protection always (e.g., during failure 1 and failure 2) switches the traffic in the egress on the single backup LSP. Local protection needs a backup LSP per link or per node being protected. In case of failure 1, traffic will be routed along backup LSP B, which is preestablished between the endpoints of the link affected by failure 1. In a similar way, backup LSP C is used during failure 2.

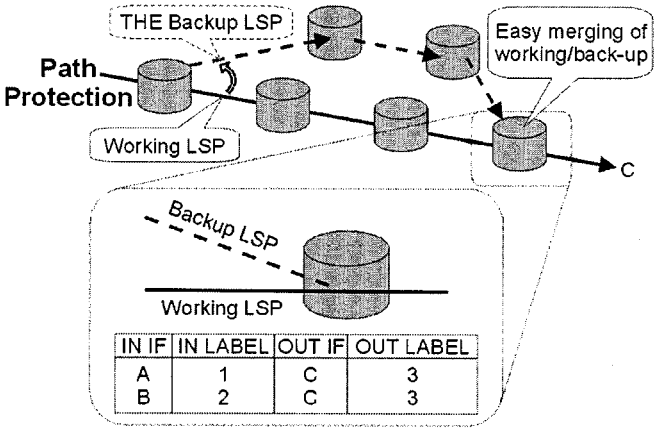

Fig. 6. Realization of merging of working and backup LSPs. Both incoming LSPs have their own entry in the label information base (LIB), and these entries target the same output interface and label. The router simply forwards any packet coming in through either working or backup LSP.

the setup of LSPs along the correct (as indicated by the routing table) path.

One of the drawbacks of MPLS rerouting is that it may suffer from similar inefficiencies as the IP routing protocols on which it is relying: e.g., rather long convergence times, temporary instabilities and loops, etc. Therefore, a new MPLS restoration scheme was developed at the Department of Information Technology (INTEC) of Ghent University: *fast topology-driven constraint-based rerouting* (FTCR); see Fig. 4. It assumes that the MPLS network runs a link-state routing protocol (e.g., OSPF or IS-IS): this means that each link is advertised to all routers in the network and that each router stores all these advertised links in its link-state database (which gives an overview of the topology). A router detecting a failure immediately knows that it has to calculate an alternative route for the LSPs leaving over the dead interface, and it may do this based on its current view of the network topology, stored in its link-state database. The router simply removes the failing equipment from the link-state database and calculates a new route from itself toward the egress LSR: this implies that the part of the LSP upstream from the failure is not rerouted. Explicitly routed setup of the LSP (i.e., specifying, in the label requests, each hop to be transited by the LSP) along this calculated alternative path is required (e.g., by means of constraint routed (CR) LDP, since other routers may not yet be aware of the failure. Later on, the IP routing protocol can continue converging/stabilizing and in the meanwhile leave the already restored LSPs alone. The principle of FTCR is illustrated with more detail in [23]–[25].

The fact that MPLS restoration sets up the LSP along the alternative path, at the moment that the failure occurs, requires only standard control plane functionality for the setup and teardown of connections. Even more, this remains true for MPλS (or any circuit-switched technology in G-MPLS).

B. Protection in MPLS

Protection in MPLS is based on a preestablished backup LSP. Such a backup LSP can span a single link or node (thus two links, in order to protect also against node failures) or a whole LSP, from ingress to egress. The former case is called *local protection*; the latter *path protection*. The upstream LSR, where the backup LSP originates, is called a *protection* (or path) *switch LSR* (PSL) and decides whether data are forwarded along the primary/working LSP or along the backup LSP. The downstream LSR, terminating the backup LSP, is called the *protection* (or path) *merge LSR* (PML) and simply merges both primary and backup LSPs into a single outgoing LSP. This MPLS protection concept is illustrated in Fig. 5.

Fig. 6 explains that merging avoids the need for a protection switch in the PML by simply forwarding any data coming in, through either the working or the backup LSP, along the outgoing LSP. Remember that IP is connectionless and thus does not require any in-order delivery of packets, even though label switched paths are introduced in MPLS-capable IP networks.

Local protection typically suffers from the fact that per link/node, a backup LSP is required for each primary LSP. Workarounds (resulting in a single backup LSP per link for all working LSPs over that link) [31] are proposed in case label

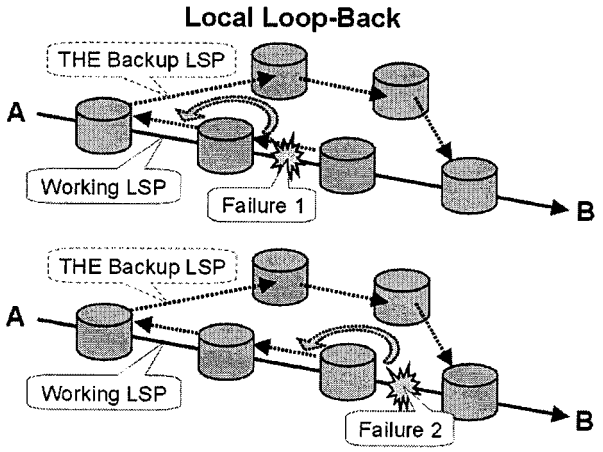

Fig. 7. The local loop-back technique combines the advantage of path protection (single backup LSP) and local protection (protection switch performed locally in an LSR adjacent to the failure). The backup LSP is routed in the opposite direction of the working LSP and continues via a disjoint route to the egress LSR. The figure clearly shows that the loop-back is performed in different LSRs (although a single backup LSP is required) under distinct failure conditions: e.g., failure 1 (top) and failure 2 (bottom).

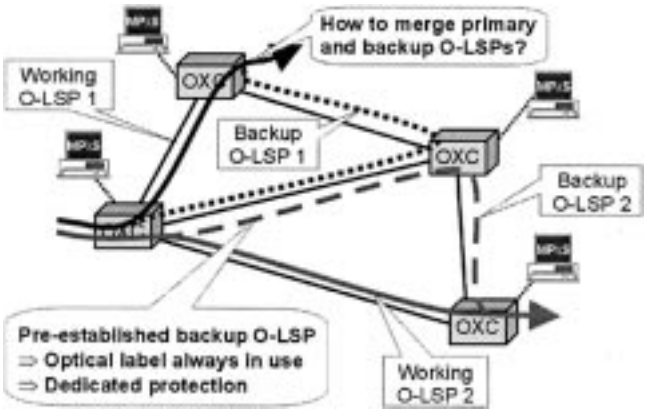

Fig. 8. Two main issues in MPλS protection. First, working and backup fixed bandwidth O-LSPs have to be merged into a single outgoing O-LSP. Secondly, MPλS protection results in dedicated protection, due to the fact that each preestablished backup O-LSP always consumes a label (or thus a wavelength), even during failure-free conditions.

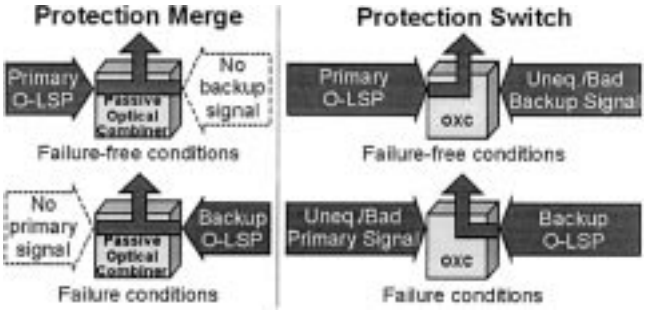

Fig. 9. A protection merge (left) can be realized by a passive optical combiner, if and only if backup and primary signals are never received simultaneously. If this condition cannot be met, a protection switch (right) is needed instead of a protection merge.

stacking is allowed and labels have a platform-wide significance. Label stacking is used to multiplex multiple LSPs into a single aggregate LSP: this is achieved by placing an additional label (e.g., shim-header) corresponding to the aggregate LSP in front of the label of the multiplexed LSPs. Platform-wide label significance means that a label space exists per LSR instead of per interface. Path protection, on the other hand, suffers from the fact that it cannot perform the protection switch locally, which requires additional signaling functionality and results in a longer interruption of the affected services or a larger amount of lost data.

The best characteristics of both protection schemes can be combined into another scheme, which we call *local loop-back* (see Fig. 7). The idea is that a single backup LSP in the opposite direction of the primary LSP allows performing of the protection switch locally. Therefore, the backup LSP consists of two parts: a reverse part, allowing the local protection switch, and a diverse part from the ingress to the egress, in order to get the protected traffic on the backup LSP through the network.

There are two main issues for protection applied to MPλS (or any circuit-switched technology in G-MPLS), as illustrated in Fig. 8.

Merging of multiple circuits into a single outgoing circuit at the same bit rate is in general not possible. Under certain conditions, specific equipment allows the implementation of a real protection merge: e.g., passively, optical combining of primary and backup signals is allowed. Fig. 9 clearly shows that this is only possible if one can assure that backup and primary signals never enter the passive optical combiner at the same time. Unfortunately, this is not always the case: one may opt to send unequipped signals over a link in order to keep the power budget on that link as constant as possible. Also, signal degradation may trigger upstream a protection switch, while the degraded primary signal is still flowing through the network. To overcome this problem, one may prefer to switch from one signal to the other one, as in classical 1+1 protection. However, this switch has to be synchronized with the status in the protection switch LSR.

Fig.8 also shows that preestablished backup LSPs result in dedicated protection, since no statistical multiplexing between circuits is allowed, as is the case in packet-switched technologies. Or in other words, a label is always required along a backup LSP, independent of whether one is dealing with a packet- or circuit-switched network, but only in a circuit-switched technology does the occupation of a label also imply the occupation of a circuit (which is considered as the capacity in such a network). This is in contrast with packet-switched technologies that allow statistical multiplexing between LSPs routed over the same link. This dedicated protection implies that MPLS protection in circuit-switched technologies may become far less efficient than in packet-switched technologies, from a capacity point of view.

III. Survivability Issues in Multilayered Networks

Our roadmap in Section I-A shows that data-centric optical networks typically consist of multiple layers, even in the simplified case of IP-MPLS directly over OTN-MPλS. This section starts with a discussion on the provisioning of recovery functionality in multilayer networks. These concepts and discussions

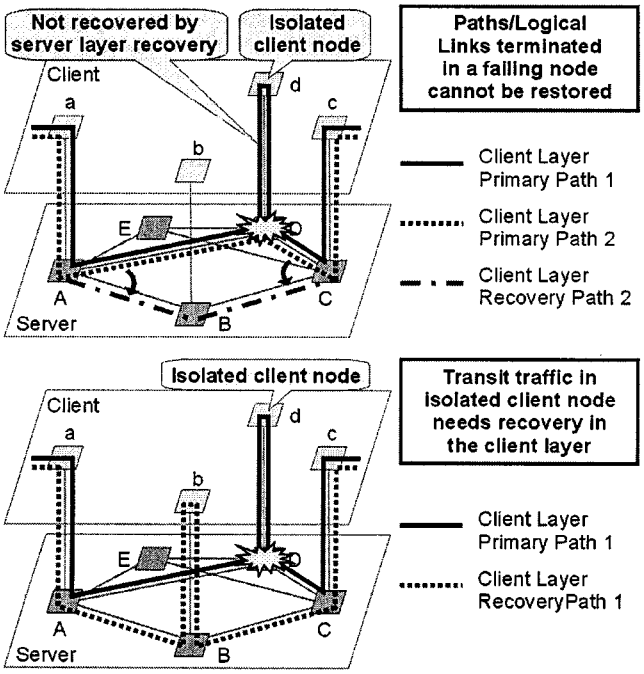

Fig. 10. Illustration of the impact of a node failure on two traffic flows between the client layer nodes "a" and "c." The top of the figure illustrates that the server layer cannot recover the first flow a-d-c (full line), because the client layer node "d" is isolated due to the failure of D, which is terminating both logical links a-d and d-c. This implies that the client layer has to recover this flow, as illustrated at the bottom of the figure. The second flow (top, dashed line) is routed over a direct logical link between nodes a and c. This logical link transits the failing node D and thus can be restored by the server layer recovery scheme.

are focused on a two-layer network but are generic and thus applicable to any multilayer network. This section ends with some survivability considerations specific to IP-MPLS directly over OTN-MPλS networks.

A. Single-Layer Survivability Strategies and Their Drawbacks

Section II gave an overview of recovery techniques applicable to MPLS or G-MPLS (e.g., MPλS) networks. However, it did not tackle the problem of in which layer to apply one of these techniques (e.g., in MPLS or in MPλS for an IP-MPLS/OTN-MPλS network). This section discusses cases where recovery is foreseen at the bottom (e.g., OTN-MPλS) or at the top (e.g., IP-MPLS) layer.

1) Survivability at the Bottom Layer: Recovery at the bottom layer has the advantage that a simple root failure has to be treated and recovery actions are performed on the coarsest granularity, resulting in the lowest number of required recovery actions. In addition, failures do not need to propagate through multiple layers before triggering any recovery action.

However, there is no recovery scheme residing in the bottom layer that can resolve any problems due to a failure in a higher layer: any layer above or the layer where the failure occurs itself has to resolve the problem. Fig.10 shows also that in the case of a node failure in the bottom layer, this layer can only recover affected traffic transiting this failing bottom layer node. The collocated higher layer node becomes isolated, and thus all traffic transiting such a higher layer node cannot be restored in the bottom layer.

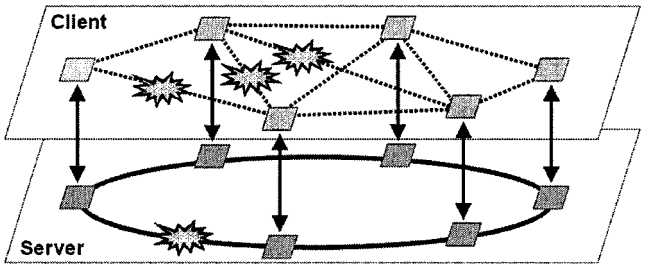

Fig. 11. A single root failure may propagate to many so-called secondary failures.

2) Survivability at the Top Layer: Another strategy is to provide the *survivability at the top layer*. The advantage of this strategy is that it can cope more easily with node or higher layer failures (see Fig. 10). A main drawback of this strategy is that it needs many recovery actions, due to the finer granularity of the flow entities in the top layer. However, treating each individual flow at the top layer allows differentiating between these flows, based on their (service) importance. In other words, the top layer may restore critical, high-priority traffic before any action is taken on low-priority flows. This is not possible in lower layers, since they switch every flow in an aggregate signal with a single action. Under certain conditions, the finer granularity may also lead to a more efficient capacity usage. First, aggregate signals, poorly filled with working traffic, have enough capacity to transport spare resources. Secondly, the finer granularity allows distributing flows over more alternative paths. However, a tradeoff exists between a better filling of the capacity of the logical links and the higher amount of higher layer equipment, when comparing this survivability at the top layer strategy with the survivability at the bottom layer strategy.

Not only the potential mismatch in granularity between the failing equipment in a lower layer and the affected entities in the top layer, requiring more recovery actions, is an issue. Also the typically complex secondary failure scenarios, as a result of a single root failure in a lower layer, can become a problem. This is illustrated in Fig. 11.

3) Slightly Different Variants: Survivability at the Lowest Detecting Layer and Survivability at the Highest Possible Layer: A slightly different variant on the survivability at the bottom layer is the *survivability at the lowest detecting layer* strategy (i.e., the lowest layer in the hierarchy able to detect the failure). This means that multiple layers deploy a recovery scheme, but still the (single) layer detecting the root failure is the only layer taking any recovery actions. With this strategy, there is no problem that the bottom layer recovery scheme will not detect a higher layer failure (because the higher layer that detects the failure will recover the affected traffic). However, this survivability at the lowest detecting layer strategy can assure that traffic transiting the failing equipment is restored, but it still suffers from the fact that it cannot restore any traffic transiting higher layer equipment isolated by a node failure. The client layer in Fig. 10 (top) deploys a recovery scheme in this strategy, but the considered traffic flow is still lost, since this client layer recovery scheme is not triggered by the node failure in the server layer. This strategy is considered as a single-layer survivability strategy, although it considers

the deployment of a recovery scheme in multiple layers. The reason is that for each failure scenario, the responsibility to recover all traffic is situated in one and only one layer (the one detecting the failure).

A slightly different variant of the survivability at the top layer strategy is the *survivability at the highest possible layer* strategy. Since not all traffic has to be injected (by the customer) at the top layer, a traffic flow is recovered in the layer in which it is injected (or, in other words, the highest possible layer for this traffic flow). For example, a data-centric optical network may also support a leased optical channel service. This strategy is also considered as a single-layer survivability strategy, although it considers a recovery scheme in multiple layers. Indeed, survivability at the highest possible layer may lead to recovery schemes in multiple layers, but never to recover the same traffic flow. Actually, for each traffic flow, a survivability at the top layer strategy is deployed (or in other words, both strategies do not differ in essence from each other).

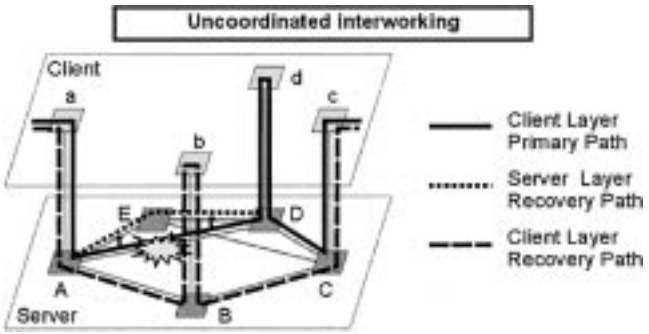

Fig. 12. The uncoordinated multilayer survivability strategy. The failure of the physical link A-D in the server layer also affects the corresponding logical link a-d in the client layer. Since recovery actions in both layers are not coordinated, they will both recover the affected traffic. The server layer reroutes all traffic on the failing link A-D through node E. The client layer restores the connection end-to-end by routing it along the path a-b-c. It is clear that in this example, recovery actions in a single layer would have been sufficient.

B. Multilayer Survivability: Concepts and Solutions

The conclusion from the previous section is that survivability both at the bottom/lowest detecting layer and at the top/highest possible layer have their pros and cons. However, it is likely that a real network will combine the advantages of both approaches. Or, more generally, the choice in which layer to recover the traffic will depend on the circumstances (e.g., the occurring failure scenario). This requires a higher flexibility than the simple rules on which the single-layer survivability strategies are based [always all recovery actions in the lowest (i.e., lowest detecting/bottom) layer or always in the highest (i.e., highest possible/top) layer].

1) Uncoordinated Approach: A first solution is to deploy a recovery scheme *in multiple layers, without any coordination*, resulting in parallel recovery actions at distinct layers. Consider for example the link failure in Fig. 12. The considered traffic flow a-c is affected and thus restored in the client layer (path a-d-c replaced by path a-b-c), while the server layer is restoring the logical link a-d (of the client layer topology) by rerouting it via node E.

The main advantage is that this solution is simple from an implementation (e.g., no standardization of coordination signals between both layers is necessary) and operational point of view. However, Fig. 12 shows the drawback of this strategy. Both recovery mechanisms occupy spare resources during the failure (i.e., the server layer along A-E-D and the client layer along a-b-c, which implies occupation of spare resources on A-B and B-C in the server layer), although one scheme occupying spare resources would be sufficient. This implies that potentially more extra traffic (i.e., unprotected preemptable traffic) is squelched (disrupted). Or even worse, consider that the server layer reroutes the logical link a-d over the path A-B-C-D instead of A-E-D; then both recovery mechanisms need spare capacity on the links A-B and B-C. If these higher layer spare resources are supported as extra traffic in the lower layer, then there is a risk that these client-layer spare resources are preempted by the recovery action in the server layer, resulting in "destructive interference." Or, rephrased, the two recovery actions taken were not able to restore the traffic, since the client layer reroutes the considered flow over the path a-b-c, which was disrupted by the server layer recovery. Reference [16] illustrates that these risks may exist in real networks: they prove that a switchover in the optical domain (e.g., for protection purposes in the optical network) may trigger traditional SDH protection.

2) Sequential Approach: A more intelligent approach, compared to the uncoordinated approach, is the sequential approach, where the responsibility for recovery is handed over to the next layer when it is clear that the current layer is not able to fulfill the recovery task. There exist mainly two approaches.

In the *bottom-up approach*, the recovery starts in the bottom/lowest detecting layer (where the failure is detected) and all traffic that cannot be restored by this layer (e.g., due to capacity shortage) will be restored by a higher layer. The advantage of this approach is that recovery actions are taken at the appropriate granularity (recovery actions on a finer granularity, in a higher layer, are only taken when necessary) and complex secondary failures are treated only when needed.

In the *top-down approach*, recovery actions are initiated in the top/highest possible layer, and only if the higher layer cannot restore all traffic are lower layer actions triggered. An advantage of this approach is that a higher layer can more easily differentiate traffic with respect to the service types, and thus it may try to restore high-priority traffic first. A drawback of this approach is that a lower layer has no easy way to detect on its own, whether a higher layer was able to restore traffic or not (an explicit signal is needed for this purpose).

The remainder of this paper assumes a bottom-up approach (since this is the most intuitive one), except when explicitly referring to the top-down approach. An example of the bottom-up approach is shown in Fig. 13. The server layer starts with attempting to restore the logical link a-d, but it fails since this logical link terminates on the failing node D. Therefore, the client layer recovery scheme is triggered to restore the considered traffic flow a-c by rerouting it over node b instead of node d.

The implementation of these escalation strategies (i.e., handing over the responsibility for recovery from one layer to the other one) is another issue. Two solutions are described here.

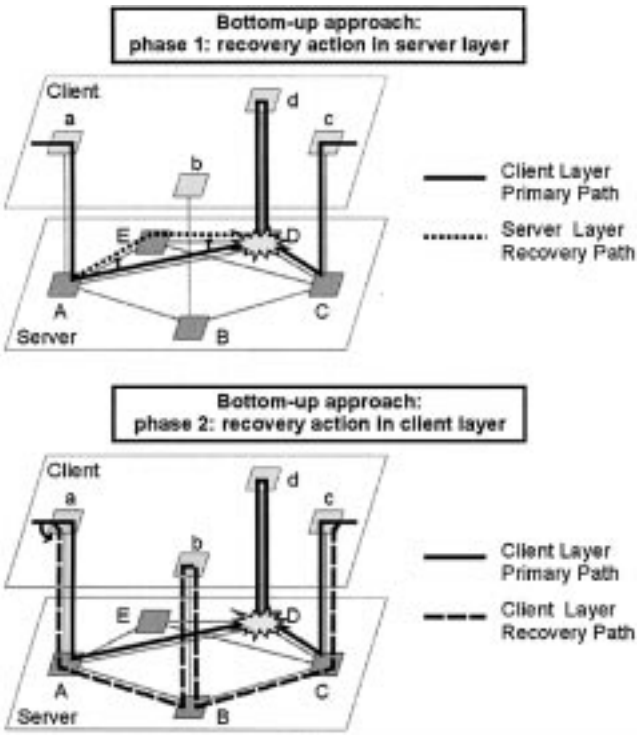

Fig. 13. The bottom-up approach. In this approach, the server layer begins trying to recover the traffic as much as possible. The logical links a-d and d-c of the client layer terminate on node D (which is failing), and thus the server layer cannot restore the traffic carried on these links. Therefore, the recovery scheme in the client layer will be triggered in one or another way. This scheme will recover the traffic transiting the isolated node d.

The first is based on a *holdoff timer*. A holdoff timer is set at the moment the server layer starts attempting to restore the traffic. If this holdoff timer goes off and (part of) the traffic is not restored, then the client layer will take over the recovery actions while the server layer ceases its attempts. The main drawback of a holdoff timer is that higher layer recovery actions are always delayed, independent of the failure scenario.

To overcome this delay, another escalation strategy is the use of a *recovery token signal* between layers. This means practically that the server layer sends the recovery token (by means of an explicit signal) to the client layer from the moment that it knows that it cannot restore traffic anymore. A disadvantage, compared to a holdoff timer interworking, is that a recovery token signal needs to be incorporated in the standardization of the interface between network layers.

A holdoff timer is probably less appropriate for a top-down approach, since the lower layer should be notified with an explicit signal whether the higher layer managed to restore the traffic or not.

3) Integrated Approach: The *integrated approach* is based on a single integrated multilayer recovery scheme. This implies that this recovery scheme has a full overview of all the network layers and that it can decide when and in which layer (or layers) to take the appropriate recovery actions. It is obvious that an integrated approach is the most flexible one. However, to profit from this high flexibility, one has to provide the necessary algorithmic intelligence/complexity. Another issue is the implementation/realization of such an integrated approach. It is unlikely to develop a single recovery scheme, controlling and having an overview of all network layers, in current overlaid networks.

TABLE I
COMPARISON AND SUMMARY OF SEVERAL PERFORMANCE PARAMETERS FOR SOME SIGNIFICANT RECOVERY STRATEGIES. THE LAST COLUMN GIVES THE TYPICALLY (BUT NOT NECESSARILY) PREFERRED VALUE FOR EACH PARAMETER

Criteria	Survivability Strategy				Preferred value
	Bottom layer	Bottom-up	Top layer	Integrated approach	
Switching granularity	Coarse	Coarse	Fine	Coarse	**Coarse**
Failure scenario	Simple	Simple	Complex	Simple	**Simple**
Recovery close to root	Yes	Yes	No	Yes	**Yes**
Capabilities, flexibility	Low	High	High	High	**High**
Failure coverage	Low	High	High	High	**High**
Coordination, mgmt	Low	High	Low	Low	**Low**
Resources	Low	High	Low	Low/High	**Low**

C. Summary and Conclusions

Section III-A discussed the shortcomings of single-layer survivability strategies. Section III-B illustrated how to overcome these shortcomings by providing survivability at multiple layers.

Table I summarizes the estimated performance, with respect to several characteristics, for some survivability strategies.

References [32] and [33] illustrate that the spare resource requirements can be reduced for the case of multilayer survivability by supporting higher layer spare resources as extra traffic in the lower layer spare resources (i.e., the *common pool* of spare resources). However, Section III-B1 explained that a proper coordination of the recovery schemes becomes absolutely necessary in such a case.

D. Specific IP-MPLS/OTN-MPλS Opportunities and Drawbacks

The goal of this section is to highlight some specific survivability opportunities and drawbacks that arise in the case of an IP-MPLS directly over OTN-MPλS network. Note that the previous sections, on generic multilayer survivability strategies, remain true for IP-MPLS/OTN-MPλS multi-layer networks: this section only provides some additional considerations, which may be taken into account when designing such an IP-MPLS/OTN-MPλS network.

Section II illustrated that MPLS is suitable to provide fast protection switching in the IP-MPLS layer. Therefore, one could opt to promote recovery in the IP-MPLS layer [i.e., promote survivability at the top/highest possible layer (e.g., [34]) or a top-down strategy], as this has some favorable properties. First, less spare resources are needed in the IP-MPLS layer, since packet switching is very suitable to *share spare*

capacity among preestablished backup paths (while keeping the advantages of fast protection switching). Secondly, *dropping low-priority* (e.g., best effort) traffic first is inherently incorporated in IP-MPLS networks if, for example, Diffserv is deployed [35].

Another opportunity relates to the integrated approach, mentioned in Section III-B3. As described in our roadmap in Section I-A, we expect that a peer-modeled data-centric optical network may become a reality in a longer term future. If this becomes true, then a *single integrated multilayer approach* would become much more feasible than in current overlaid networks, due to the single integrated control plane of a peer-modeled network.

Finally, the automation of the lightpath setup/teardown process in an ASON does not require one to stick with a fixed logical (IP-MPLS) topology and capacity. This opens opportunities for the reoptimization of the logical topology during a failure condition. Even more, biconnected logical (IP-MPLS) topology is no longer an absolute necessity to survive any single failure. For example, if a router would fail (potentially resulting in a disconnected IP-MPLS network), an automatic reconfiguration of the logical IP-MPLS topology [instead of traditional rerouting (i.e., protection/restoration) of traffic] would restore the connectivity of the IP-MPLS network.

A main drawback of current IP-MPLS network is that failure detection is based on the periodic exchange of Hello messages between adjacent routers. If no Hellos are received through an interface, then the only conclusion can be that the opposite side of the interface is unreachable or, in other words, that each packet sent through the interface is sent into a black hole. But this detection scheme does not allow one to differentiate between a router failure (meaning that the router at the opposite side of the link is dead) in the IP-MPLS layer itself and a failing logical link in the IP-MPLS layer as a result of a failure in the OTN-MPλS layer. This implies that the survivability at the lowest detecting layer is impossible in an IP-MPLS/OTN-MPλS network.

Another concern of this Hello message detection scheme is the detection time. Current IP routers send a Hello message each 10 s, and a defect is declared after the loss of four Hello messages (resulting in a typical detection time of 40 s, which gives the OTN a lot of time to fix the problem without the IP layer even noticing) [36]. However, driving this periodicity to the order of milliseconds becomes reasonable in IP-MPLS/OTN-MPλS, due to the huge capacity (e.g., 10 Gb/s) of a logical link, resulting in an insignificant bandwidth overhead for the Hello messages.

IV. Case Studies on Survivability in IP-MPLS Directly Over OTN-MPλS Networks

The goal of this section is to present case study results that deal with survivability in data-centric optical networks. First, typical network scenarios are described. Then two studies are presented that may help in deciding in which layer (IP-MPLS or OTN-MPλS) to provide survivability functionality. The first study compares the cost of MPLS protection whether deployed in the electrical IP-MPLS or optical OTN-MPλS layer. The second one studies the influence of protection switching, and its timing, on TCP behavior (which is typical for data traffic). This section ends with the design of a sample network that may or may not take into account IP-MPLS router failures.

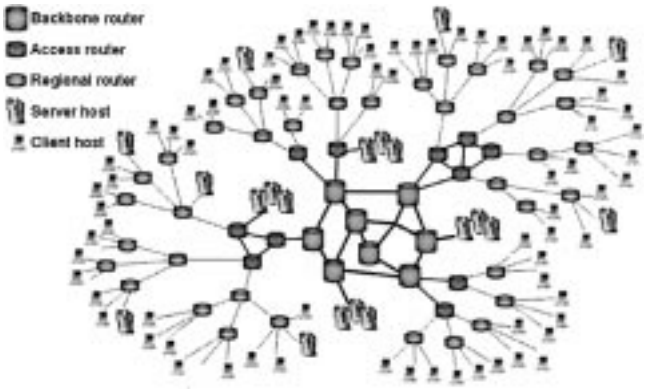

Fig. 14. Typical IP-MPLS network topology (backbone+access part).

A. Typical Network Scenarios

A typical IP-MPLS network consists of a meshed core network containing a few dozens of backbone IP-MPLS routers. Attached to those backbone routers are regional networks that concentrate the traffic from the access part of the network into the core part. While the core part of the network has a meshed structure, the structure of the access part of the network could be described as a tree structure, as illustrated in Fig. 14. Also attached to the IP-MPLS network are large server farms, containing the data for, e.g., video-on-demand or Web-based services. They are one of the reasons of the highly asymmetric character of IP traffic (e.g., video-on-demand: small customer request stream in the upstream direction, large video-data stream in the downstream direction) [37].

In the IP-MPLS directly over OTN-MPλS scenario considered in this paper, the logical (backbone) IP-MPLS links are directly supported by optical paths in the OTN-MPλS layer. However, various routing options still exist, especially in the backbone part of the network. Some operators will probably have a single-hop IP-MPLS core network where traffic is routed through only two backbone routers: one through which it enters the backbone network and one through which it leaves the backbone. This implies, of course, that the backbone part of the network is a full mesh on the logical IP-MPLS level. Other operators might have a multihop network in which the IP traffic traverses several logical links (hops) before it leaves the backbone. Since LSPs will typically start and/or terminate somewhere in the access part of the network (or even at a host), most LSPs will pass through multiple routers (even in the case of a single-hop logical core network).

B. Recovery at MPLS and/or at the MPλS Layer?

An important issue in this paper is in which layer to provide a recovery scheme. The goal of this section is to present some quantitative study results, which may help to answer this question.

A first study investigates the amount of required spare resources relative to the amount of working resources. The previous section and Section II-.B explained that MPLS protec-

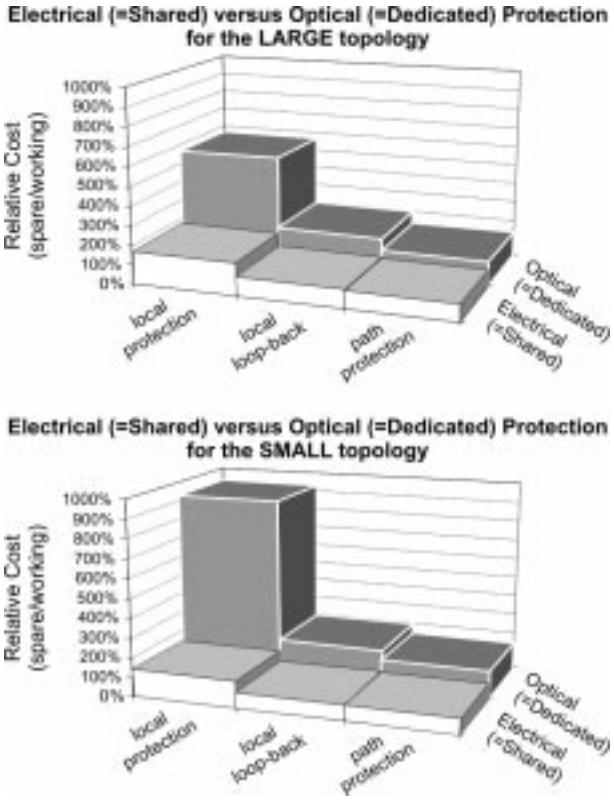

Fig. 15. Comparison of electrical and optical MPLS protection, from a transmission line capacity or cost (i.e., capacity multiplied by distance) point of view. The presented study confirms that 1) optical protection is more expensive (due to dedication of spare resources) and 2) this is most drastic for local protection, which 3) is the most expensive scheme anyway.

tion results in shared protection when applied at the electrical MPLS layer and in dedicated protection at the optical layer. The goal of our study is to investigate the significance of this effect by comparing the results for both cases. Fig. 15 makes such a comparison between both relative values, for all MPLS protection schemes described in Section II-B, for two topologies. The large topology contains 44 nodes, interconnected by 57 links, resulting in an average nodal degree of 2.59. The small topology contains 30 nodes, interconnected by 36 links, resulting in an average nodal degree of 2.4. The values presented in the charts are an average over ten randomly generated traffic matrices. The routing strategy is as follows. First, the working route is calculated based on a Dijkstra shortest path algorithm. Subsequently, the shortest node-disjoint route is computed for path protection and local loop-back. It can happen that such a route is not found, which implies that traffic is lost (or not protected) during a failure by both schemes. Local Protection is based on backup paths spanning two links, in order to be able to protect also against node failures. There is only one exception: a backup path is also spanned over the last link of each connection, since it would make no sense to send the traffic one hop behind the termination node, in case the last link would fail.

Fig.15 indeed confirms our expectation that dedicated MPLS protection (thus in the optical layer) is more expensive than shared MPLS protection (in the electrical MPLS layer). More important is that these charts show that the difference is severe for local protection. This result can be sensed as follows. Sharing between two (or more) backup paths using the same

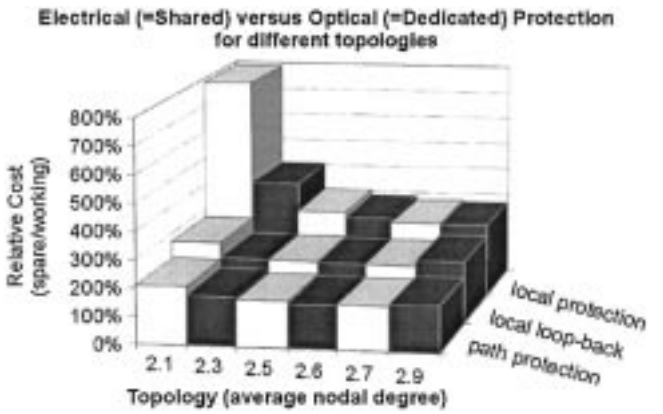

Fig. 16. Illustration of the increase in required spare resources, due to dedication, for a set of topologies with different degrees of meshedness. The figure illustrates that this is most drastic for local protection, especially on sparse topologies (which are unfortunately typical for optical transport networks).

resource is only possible if the two corresponding working segments (a segment is a path in the case of path protection or local loop-back, one link in the case of link protection, and two links in case of node protection) do not overlap. In the case of local protection, these working segments are in general shorter than for path protection or local loop-back (one or two links versus a complete path), implying a smaller probability of working segments' overlapping and hence a higher probability that sharing between the two backup paths is indeed allowed. Hence, the relative difference between dedicated and shared protection in terms of capacity requirements will be more substantial for local protection than for path protection or local loop-back.

Fig. 15 also reveals that (in the case of local protection) the topology has a significant impact, and more precisely that the topology with the smallest nodal degree suffers the most from this dedication. This result can be understood intuitively as follows. If a topology becomes sparser, backup LSPs for adjacent failure scenarios (e.g., two adjacent links in case of link protection) tend to become longer and more overlapping (for instance, think about the extreme case of a ring topology to sense this). This explains why the penalty of dedication is severe in sparse networks and less in dense networks. These observations are confirmed by the study in Fig. 16 investigating the impact of the (nodal degree of a) topology on the relative cost increase due to the dedication of MPLS protection in the optical domain. As in the exercise of Fig. 15, for each topology an average of the transmission-line cost was calculated over ten randomly generated traffic matrices, and the same routing strategy was assumed. The topologies were derived from an existing 20-node network by removing some links in order to reduce the connectivity (or the average nodal degree). The conclusion is that fast MPLS protection in the electrical MPLS layer is cheaper than similar schemes in the optical transport network and that the cost increase for local protection in the optical layer could be very severe, due to the typical sparse topologies of transport networks. More detailed spare resource dimensioning results for MPLS recovery schemes can be found in [24] and [23].

Another issue is whether the dominant data traffic (typically based on TCP) prefers fast protection switching. Assume that one wants to profit from the advantages of fast protection

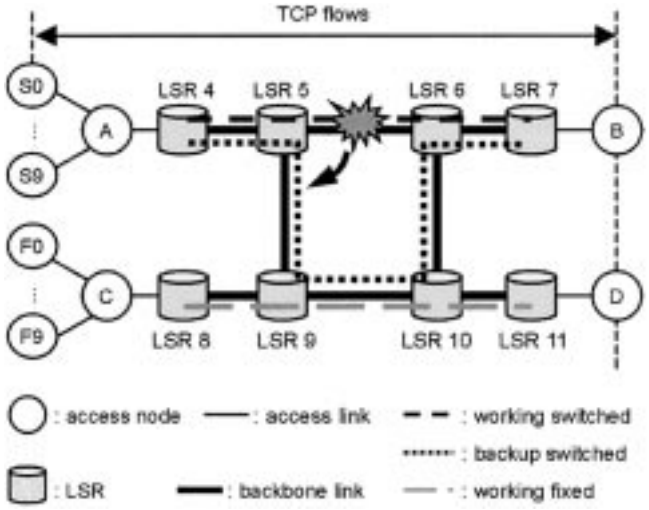

Fig. 17. Simulation topology investigating effect of protection switching on TCP.

switching in the electrical MPLS layer. Then, there may be a risk that switching a large amount of traffic (e.g., a complete 10-Gb/s line) immediately (i.e., before the TCP mechanism gets the chance to slow down) would drastically impact other flows in the network. Indeed, as TCP is reactive in nature, not only the flows being switched to an alternative backup path will be affected but also the other flows [already present on (parts of) the backup path]. To gain a better understanding of these kind of interactions, and the role of the exact timing of the protection switch, a simulation study was carried out.

The setup of the simulation is depicted in Fig. 17. We consider a backbone network of LSRs to which we connect access nodes via links having a bandwidth of 90% of the backbone links. In this network, we set up two categories of flows. The "switched flows" will follow the path crossing LSRs 4, 5, 6, and 7 when there is no link failure; upon the failure of link 5-6, a protection switch will be carried out at LSR 5 and the followed path will be 4-5-9-10-6-7, as indicated by the dotted line in Fig. 17. The other category, the "fixed flows," will always use the path over LSRs 8, 9, 10, and 11. The simulation scenario consists of three periods of 5 s: during the first and third, all links will be up, whereas during the second period link 5-6 will fail. To investigate the influence of timing, the protection switch will be performed "manually" exactly δ s after the occurrence of the link failure.

From a qualitative point of view, the influence of δ can be easily predicted. If δ is set to zero, the switched flows will join the fixed ones at LSR 9 at a time when they are both sending at a quite high rate (limited only by the bandwidth of the access links). This will result in an almost immediate buffer overflow at LSR 9, causing a burst of a fairly high number of losses, afflicted on both flow categories. Introducing a small delay (δ strictly positive) will inflict losses during that period of δ on the switched flows only, thereby forcing them to back off (compare to TCP window size reduction in response to losses) before being switched to the alternative path. As a result, the immediate buffer overflow at LSR 9 will be avoided and the fixed flows will be approached more "gently": a buffer overflow at LSR 9 will occur at a later time and will cause fewer losses compared to

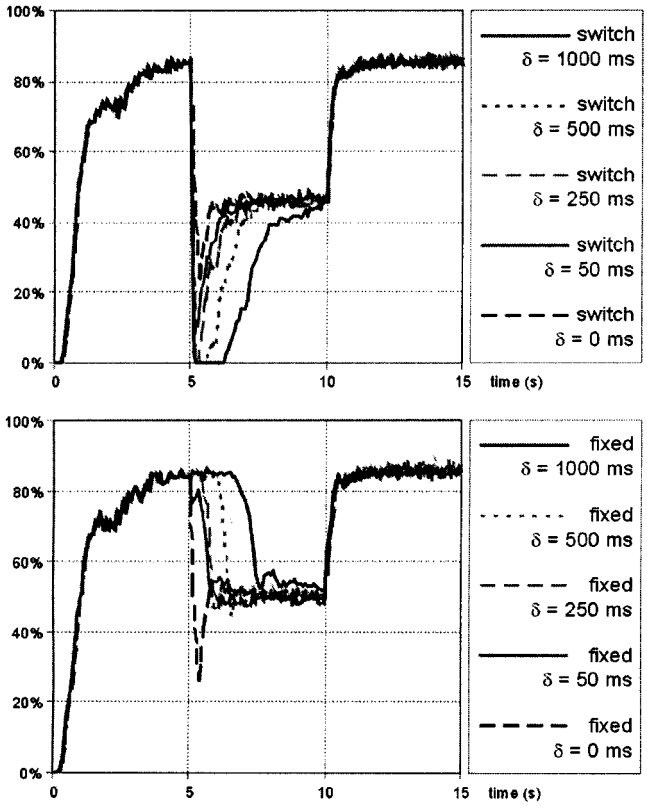

Fig. 18. TCP goodput evolution over time for different values of the protection switching delay δ. The graph at the top shows the goodput attained by the whole of the switched flows, whereas the bottom contains the evolution for the fixed flows. The goodput is expressed in percentage of backbone link bandwidth; it was measured with a resolution of 10 ms.

the $\delta = 0$ case. In Fig. 18, the evolution of goodput over time is depicted. There we clearly see the heavy impact (i.e., serious drop in goodput) of the immediate buffer overflow for $\delta = 0$ on the fixed flows.

To decide what delay δ results in the "best" behavior from a quantitative point of view, we decided to use TCP goodput as a decision criterion. Indeed, goodput is what an end user cares about: it is the amount of data successfully transported end-to-end during a certain time interval (expressed in, e.g., bytes/s). We ran simulations using random start times for the TCP sources and randomly generated propagation delays for the first access links [to introduce diverse round-trip times (RTTs) for different source-destination pairs]. For each of the thus created 150 random cases, we ran simulations for five different values of δ (0, 50, 250, 500, and 1000 ms) tracing TCP goodput. We compared the different values of δ by plotting the histogram of the ratio $f(\delta) = \text{Good}(\delta)/\text{Good}(0)$, where $\text{Good}(\delta)$ is the total goodput—attained by the whole of fixed and switched flows—during the first 1.5 s after the link failure for delay δ (we chose 1.5 s, as we intended to focus on the smaller delays and this is the relevant period for those cases). These histograms (and corresponding normal fits) are depicted in Fig. 19. The graph shows that on average, all cases of δ result in a better overall goodput than having no delay at all ($\delta = 0$).

The measurements of Fig. 19 are interpreted and summarized in Table II.

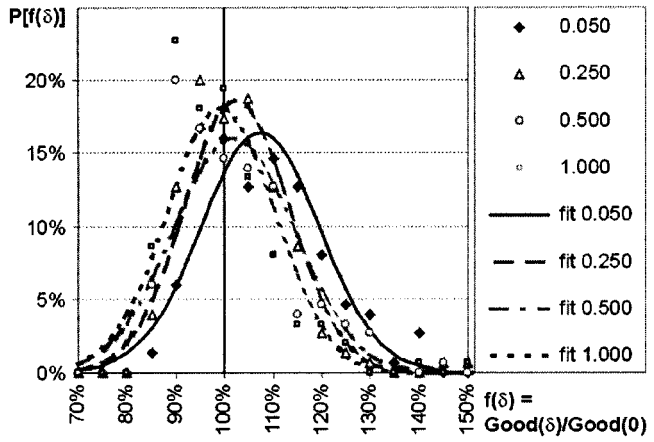

Fig. 19. Histograms (with a resolution of 5%) and normal fits for relative amount of goodput. A marker at (x, y) for a particular δ means that y% of the simulation results had $f(\delta)$ within $[x, x + 5\%]$.

TABLE II
COMPARISON OF DIFFERENT PROTECTION SWITCH DELAYS. THE LEFT COLUMN REPRESENTS THE X-VALUE CORRESPONDING TO THE AVERAGE OF $F(\delta)$, I.E., THE PEAK OF THE NORMAL FIT IN FIG. 19. THE SECOND COLUMN INDICATES THE PERCENTAGE OF SIMULATION RESULTS WHERE $F(\delta) < 100\%$. THE RIGHTMOST COLUMN GIVES THE NUMBER OF SIMULATION RESULTS WHERE GOOD(δ) WAS MAXIMAL (I.E., COMPARED TO OTHER DELAYS)

δ	average better goodput than case δ=0	% of random cases where δ is worse than δ=0	% of random cases where delay δ is best
0.000	0.00%	0.00%	20.00%
0.050	9.85%	24.00%	64.67%
0.250	4.99%	36.67%	9.33%
0.500	4.39%	42.67%	5.33%
1.000	1.75%	49.33%	0.67%

An important observation that can be drawn from these simulation results is that the time it takes for the interacting TCP flows to stabilize after the protection switch is on the order of a second or more (see Fig. 18). It can be concluded that pushing fast protection switching to the limit (i.e., extremely fast) may not be the best thing to do. However, deciding upon the "best" time to perform the protection switch is not easy. It depends at least on the link load (in the case presented above, when all links are up, backbone links are loaded for maximum 90% due to the limits in the access part, but a protection switch results in a sudden load of almost 180%), the RTT experienced by the TCP sources (larger RTT means slower response to topology changes), and the number of concurrent TCP flows (larger number results in faster stabilization, up to a certain limit).

However, the results presented above seem to indicate that from a practical point of view, it is not harmful to have fast protection (order of tens of milliseconds) for TCP traffic. This conclusion is probably even more true if we believe that backbone links carry a vast amount of concurrent TCP flows (compare to faster stabilization than smaller number of flows, and therefore optimal delay shifts toward $\delta = 0$) and/or are fairly underloaded. Indeed, when backbone links do not form the bottleneck for TCP flows, interaction between switched and fixed flows will be limited. Other simulations showed that in this latter case (e.g., for an access link bandwidth that is 60% of the backbone bandwidth), the optimal protection switch delay clearly shifts to lower values (toward $\delta = 0$). The simulations carried out so far seem to indicate that only if the timescale of protection switching is well below 50 ms may TCP effects call for a stop to the efforts to minimize it. All this, however, does not imply that extremely fast protection switching is a must for TCP: the differences in goodput for delays in the range 0–250 ms do not differ all that much, especially when the number of TCP flows is large.

The simulation discussed above considered fast protection at the MPLS layer. However, if fast protection is offered by lower layers (e.g., MPλS), we are in an altogether different situation. Indeed, in that case, we will have no interaction between competing TCP flows (as we assume that the capacity for protection is reserved and is fully available from the very instant the protection switch is carried out); clearly, dynamic behavior of TCP in response to packet losses will still occur. In this case, the intuitively clear conclusion we have drawn from a first series of simulations is: the faster the protection switch at the optical layer is performed, the better (from a TCP goodput point of view). The simulations performed for this case had a link going down for a certain amount of time δ, without any protection actions taken at the MPLS level. For 140 random cases (random RTTs, etc., as before) and δ in {0, 5, 10, 20, 30, 40, 50, 250, 500, 1000 ms}, we saw that in 94% of the cases, $\delta = 0$ was the best (only packets in transit on the failing link are lost); in the remaining 6% of the cases, $\delta = 5$ ms was the best (which is due to details in dynamic TCP behavior in some rather peculiar cases). Thus, the avoidance of TCP interactions is an advantage of protection at the MPλS layer and means that even extremely fast protection switching at that layer does not seem to pose any problem (at least from a TCP point of view).

We conclude this section by saying that from a capacity point of view, protection in the MPLS layer is preferable to MPλS protection. However, one has to be careful when performing fast protection switching in the MPLS layer, since TCP may behave in such a way that its goodput slightly reduces when switching too fast. Thus, this section illustrates that such a decision is far from straightforward.

C. Case Study: Design of a Multilayer Survivable MPLS/OTN Network

The concept of survivability in a multilayer network is illustrated here with an example. The network under study is an MPLS-over-OTN network [38]. Both layer networks are shown in Fig. 20. The MPLS layer contains 16 routers, connected by 33 logical links. Attached to the routers of the major cities are servers that contain the application data (e.g., video data for the video-on-demand service). The topology resembles a multiple star topology, with the heart of each star in a router connected to a large server (farm). The OTN layer is made up of 14 OXCs and 29 links in a mesh topology. Both topologies are biconnected.

The starting point of the design is the IP traffic matrix, which combines the demands of the various IP services (voice-over-IP, video-on-demand, Web-based services, e-mail, etc.). This matrix is asymmetric, as some of the routers are connected to server farms and thus generate large amounts of traffic, which

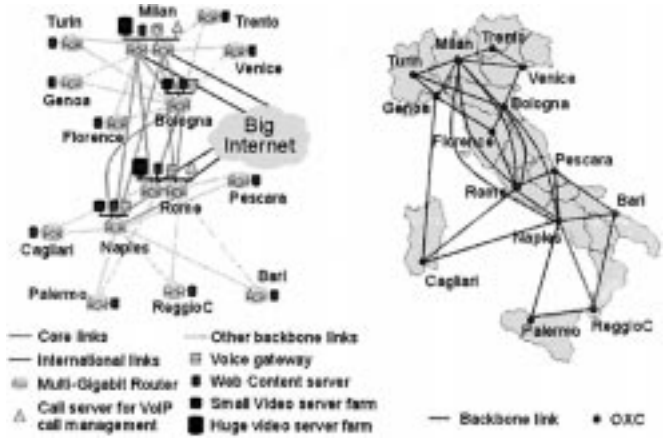

Fig. 20. (a) MPLS topology and (b) OTN topology.

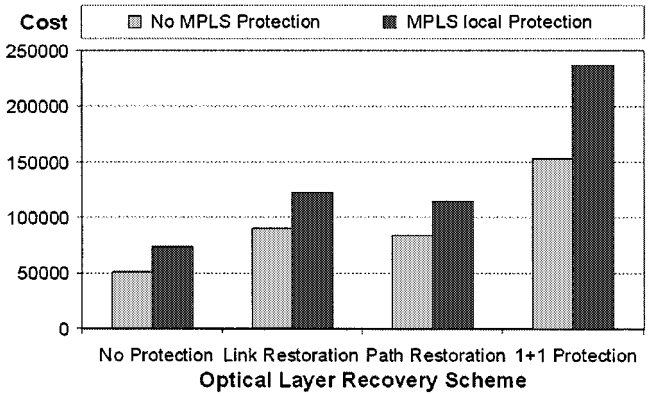

Fig. 21. The overall network cost for different resilience strategies. Per recovery technique in the optical layer, two values are given: one for the case with and one for the case without local MPLS protection against router failures.

is downloaded by users scattered all over the country. Based on the IP traffic demand and the MPLS topology, the MPLS layer is dimensioned, using an MPLS-based planning tool. It routes the unidirectional IP traffic along the shortest path between its source and destination. Because the individual unidirectional flows are routed one at a time, it is possible that both directions between the same ⟨source, destination⟩ pair are routed along different paths (with equal lengths). To provide recovery for LSR failures (or any other failure isolating an LSR), the network can be dimensioned for MPLS local protection (see Section II-B).

The MPLS dimensioning tool gives as output the routing of the traffic (on each link) and thus the capacity needed on the MPLS links. These are fed into the OTN planning tool together with the OTN topology. The maximum capacity on both directions of a logical link is considered as the capacity needed on that link, or in other words, the number of bidirectional lightpaths to be set up between two LSRs.

Line systems of 32 wavelengths were assumed, with each wavelength carrying an STM-16 signal. The routing in the OTN layer starts from an initial shortest path routing and tries to remove inefficiently used line systems by rerouting the traffic along other line systems that have enough unused capacity left. The tool can calculate the spare resources needed for different recovery schemes: no protection, link or path restoration, and 1+1 protection [39]. In our design, the OTN layer was chosen to provide resilience against expected failures (this includes single link and node failures). However, as described in Section III-A1), a recovery scheme in the OTN layer alone does not suffice to provide resilience against MPLS router failures (or any other failure isolating a router). An appropriate recovery scheme in the MPLS layer (e.g., MPLS local protection) is needed. This will result in an increase of the overall cost, because extra capacity in the OTN is needed to support the spare resources of the MPLS layer. Fig. 21 shows a comparison in terms of cost between the various possible recovery schemes in the OTN, with and without the use of MPLS local protection in the MPLS layer.

The cost is modeled as the sum of the number of wavelengths needed on the various links multiplied by the link length. Also important here is the assumption that all OXCs are able to perform wavelength conversion.

A first conclusion that can be drawn from these results is that the use of 1+1 protection in the OTN layer leads to the most expensive solution, 1.7 to 1.8 times more expensive than restoration (in the case where no MPLS protection mechanisms are used). Path restoration is in this case the cheapest solution. A second result is that the introduction of MPLS local protection has a serious impact on the overall cost. On average, the network cost increases with a factor of 1.4 due to its use. In this case, the extra cost of 1+1 protection compared to restoration is even higher: 1.9 to 2 times more expensive. Again, path restoration is the cheapest solution. Of course, the network is now also protected against MPLS router (isolating) failures, which was not true in the former case. However, part of the cost increase can be explained by the fact that spare resources are now needed in both layer networks. This results in what is called redundant or double protection: spare resources in the OTN layer also protect spare resources from the MPLS layer, which is superfluous. This can be avoided by supporting the MPLS spare resources as unprotected traffic in the OTN. Even better results can be obtained by adopting a multilayer survivability strategy based on the common pool concept [32], [33]. The basic idea behind this concept is to support higher layer spare resources as unprotected preemptable traffic in the lower layer network.

V. Conclusion

A roadmap has been outlined in this paper, showing how current core networks will evolve from a rather complex IP/ATM/SDH/WDM toward a simplified IP-directly-over-OTN paradigm. In particular, the survivability features of such data-centric optical networks have been investigated. Special attention has been paid to the application of MPLS recovery techniques.

Since data-centric optical networks contain at least an IP-MPLS layer and an optical layer, one of the main questions to be answered was: “In which layer should survivability features be provided?” It was shown that each layer has its pros and cons. Therefore, a likely solution seems to be to provide survivability at multiple layers in order to combine the advantages of these layers. However, in order to avoid

inefficiencies or conflicts between these layers, the recovery actions of these layers may require coordination. Therefore, in addition to the uncoordinated approach, a sequential (e.g., by means of a holdoff timer or recovery token) and an integrated approach have been proposed.

Finally, some case studies illustrated the relevance of those multilayer survivability issues. One of the conclusions was that MPLS protection allows fast recovery of traffic at the electrical MPLS level, and even more that this is typically cheaper than MPλS protection, but protection switching at the MPLS level may have a negative impact on TCP goodput during a rather long period [on the order of a (few) second(s)] after the failure and the protection switch. Another case study illustrated that protecting against MPLS router failures while trying to recover as much traffic as possible in the OTN, without appropriate precautions, may have a significant negative impact on the overall network cost.

REFERENCES

[1] Evaluating the size of the Internet [Online]. Available: http://www.netsizer.com

[2] The Emerging Digital Economy (1998, Apr.). [Online]. Available: http://www.ecommerce.gov/emerging.htm

[3] *Evolution of Information and Communication and Its Impact on Research Activities*, Aug. 30, 1999. EISI-WAY.

[4] The European Information Technology Observatory [Online]. Available: http://www.eito.com

[5] P. Green, "Progress in optical networking," *IEEE Commun. Mag.*, vol. 39, pp. 54–61, Jan. 2001.

[6] K. Struyve *et al.*, "Application, design and evolution of WDM in GTS's pan-european transport network," *IEEE Commun. Mag.*, vol. 38, pp. 114–121, Mar. 2000.

[7] N. Ghani, "Lambda-labeling: A framework for IP-over-WDM using MPLS," *Opt. Networks Mag.*, vol. 1, no. 2, pp. 45–58, Apr. 2000.

[8] J. Anderson, J. S. Manchester, A. Rodrigyez-Moral, and M. Veeraghavan. (1999, Jan.–Mar.) Protocols and architectures for IP optical networking. *Bell Labs Tech. J.* [Online], pp. 105–124. Available: http://www.lucent.com/minds/techjournal/jan-mar1999/pdf/paper06.pdf

[9] K. Shiomoto, M. Uga, M. Omotani, S. Shimizu, and T. Chimaru, "Scalable multi-QoS IP+ATM switch router architecture," *IEEE Commun. Mag.*, vol. 38, pp. 86–92, Dec. 2000.

[10] "Multi-protocol label switching," *IEEE Commun. Mag.*, vol. 37, no. 12, Dec. 1999.

[11] R. Callon, P. Doolan, N. Feldman, A. Fredette, G. Swallow, and A. Viswanathan, A framework for multi-protocol label switching, in IETF Internet Draft $\langle$draft-ietf-mpls-framework-05.txt$\rangle$, Sept. 1999. work in progress.

[12] G. Armitage. (2000, Jan.) MPLS: The magic behind the myths. *IEEE Commun. Mag.* [Online], pp. 124–131

[13] G. Swallow, "MPLS advantages for traffic engineering," *IEEE Commun. Mag.*, vol. 37, pp. 54–57, Dec. 1999.

[14] L. Andersson, P. Doolan, N. Feldman, A. Fredette, and B. Thomas. (2001, Jan.) LDP Specification. [Online]. Available: http://www.ietf.org/rfc/rfc3036.txt

[15] Internet Draft <draft-ietf-mpls-rsvp-lsp-tunnel-08.txt> (2001, Feb.). [Online]. Available: http://www.ietf.org/internet-drafts/draft-ietf-mpls-rsvp-lsp-tunnel-08.txt

[16] N. Wauters, G. Ocakoglu, K. Struyve, and P. F. Fonseca, "Survivability in a new pan-european carriers'-carrier network based on WDM and SDH technology: Current implementation and future requirements," *IEEE Commun. Mag.*, vol. 37, pp. 63–69, Aug. 1999.

[17] Deliverable 2 of the Eurescom Project P918 (2000, Sept.). [Online]. Available: http://www.eurescom.de/public/projectresults/results.asp

[18] A. Neukermans and R. Ramaswami, "MEMS technology for optical networking applications," *IEEE Commun. Mag.*, vol. 39, pp. 62–69, Jan. 2001.

[19] *User Network Interface (UNI) 1.0 Signaling Specification*, June 2001.

[20] *Architecture for the Automatic Switched Transport Network*, ITU-T G.astn v.0.3.

[21] IETF Internet Draft <draft-awduche-mpls-te-optical-03.txt> (2001, Apr.). [Online]. Available: http://www.watersprings.org/pub/id/draft-awduche-mpls-te-optical-03.txt

[22] IETF Internet Draft <draft-ietf-mpls-lsp-hierarchy-02.txt> (2001, Feb.). [Online]. Available: http://www.watersprings.org/links/mlr/id/draft-ietf-mpls-lsp-hierarchy-02.txt

[23] D. Colle, A. Groebbens, P. Van Heuven, S. De Maesschalck, M. Pickavet, and P. Demeester, "Porting MPLS-recovery techniques to the MPλS paradigm," *Opt. Networks Mag. (Special Issue on Protection and Survivability)*, vol. 2, no. 4, July/Aug. 2001. .

[24] D. Colle *et al.*, "MPLS recovery mechanisms for IP-over-WDM networks," *Photon. Network Commun. Mag. (Special Issue on IP over WDM and Optical Packet Switching)*, vol. 3, no. 1/2, pp. 23–40, Jan.–June 2001. .

[25] P. Van Heuven, S. De Maesschalck, D. Colle, S. Van den Berghe, M. Pickavet, and P. Demeester, "Recovery in IP based networks using MPLS," in *Proc. IEEE Workshop IP-Oriented Operations & Management (IPOM'2000)*, Cracow, Poland, Sept. 2000, pp. 70–78.

[26] IETF Internet Draft <draft-haskin-mpls-fast-reroute-05.txt> (2000, Nov.). [Online]. Available: http://www.watersprings.org/links/mlr/id/draft-haskin-mpls-fast-reroute-05.txt

[27] IETF Internet Draft <draft-makam-mpls-recovery-frmwrk-01.txt> (2000, July). [Online]. Available: http://www.watersprings.org/links/mlr/id/draft-makam-mpls-recovery-frmwrk-01.txt

[28] IETF Internet Draft <draft-makam-mpls-protection-00.txt> (1999, Oct.). [Online]. Available: http://www.watersprings.org/links/mlr/id/draft-makam-mpls-protection-00.txt

[29] IETF Internet Draft <draft-chang-mpls-path-protection-02.txt> (2000, Nov.). [Online]. Available: http://www.watersprings.org/pub/id/draft-chang-mpls-path-protection-02.txt

[30] Y. Ye, S. Dixit, and M. Ali, "On joint protection/restoration in IP-centric DWDM-based optical transport networks," *IEEE Commun. Mag.*, vol. 37, no. 8, pp. 174–183, Aug. 1999.

[31] IETF Internet Draft <draft-swallow-rsvp-bypass-label-01.txt> (2000, Nov.). [Online]. Available: http://www.watersprings.org/pub/id/draft-swallow-rsvp-bypass-label-01.txt

[32] M. Gryseels, K. Struyve, M. Pickavet, and P. Demeester, "Common pool survivability for meshed SDH-based ATM networks," in *Proc. Int. Symp. Broadband European Networks (SYBEN'98)*, Zurich, Switzerland, May 1998, pp. 267–278.

[33] P. Demeester *et al.*, "Resilience in multi-layer networks," *IEEE Commun. Mag.*, vol. 37, pp. 70–76, Aug. 1999.

[34] G. Heron, "Level 3: MPLS over DWDM," presented at the IP Over DWDM Conf., Paris, France, Nov. 27–30, 2000.

[35] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. (1998, Dec.) An architecture for differentiated services. [Online]. Available: http://www.ietf.org/rfc/rfc2475.txt

[36] J. Moy. (1991, July) OSPF Version 2. [Online]. Available: http://www.ietf.org/rfc/rfc1247.txt

[37] M. Listani and R. Sabella, "Architectural and technological issues for future optical internet networks," *IEEE Commun. Mag.*, vol. 38, pp. 82–92, Sept. 2000.

[38] D. Colle *et al.*, "Envisaging next-generation data-centric optical networks," in DRCN2001 Conf., submitted for publication.

[39] P. Arijs, B. Van Caenegem, P. Demeeester, P. Lagasse, and W. Van Parijs, "Design of ring and mesh based WDM transport networks," *Opt. Networks Mag.*, vol. 1, no. 3, pp. 25–40, July 2000.

Didier Colle graduated as an electrotechnical engineer (option communication technology) from Ghent University, Belgium, in 1997.

Since 1997, he has been with the Department of Information Technology, Ghent University, in the Broadband Communications Network group. His experience lies in the field of network planning: design of SDH networks and evaluation of a commercial available SDH planning tool. Currently, he is working on the planning of IP/x and (G-)MPLS networks. He is strongly interested in the implications on the planning results due to the limitation of the equipment (e.g., reduced flexibility of ADMs). He is currently actively involved in two European projects: the IST-project LION and the COST-action 266.

Sophie De Maesschalck received the M.Sc. degree in electrotechnical engineering (option communication technology) from Ghent University, Belgium, in 1998, where she currently is pursuing the Ph.D. degree.

Since 1998, she has been a Research Assistant at the Department of Information Technology, Ghent University. Her main research interests include the design, planning, and evaluation of IP networks and optical transport networks. She was involved in the ACTS project SONATA and is currently working on the IST project LION.

Chris Develder received the M.Sc. degree in computer science engineering from Ghent University, Belgium, in 1999, where he currently is pursuing the Ph.D. degree.

In October 1999, he joined the Department of Information Technology (Broadband Communication Networks Group), Ghent University, as a Researcher for the Fund for Scientific Research of Flanders. There, he is working in the field of network design and planning, mainly focussing on optical packet switched networks (IP/x, (G)MPLS-based, etc.). As such, he is involved in two European IST-projects: LION and DAVID.

Pim Van Heuven graduated in computer science from Ghent University, Belgium, in 1998, where he currently is pursuing the Ph.D. degree.

His thesis ("Computer Busses and Caches in Future Processors") examined the benefits of prefetching in future processors by means of simulation. In August 1998, he joined the Broadband Communications Networks Group (IBCN). In January 1999, he received an IWT scholarship. His research topics include MPLS, QoS, and traffic engineering. He was involved in the ACTS Ithaci project and is currently working on the IST Tequila project.

Adelbert Groebbens graduated as a computer science engineer (option communication technology) from the University of Ghent, Belgium, in 2000.

In October 2000, he received an FWO aspirant scholarship. Since then, he has been a Researcher in the Broadband Communications Networks Group (IBCN), Department of Information Technology (INTEC), Ghent University. His interests include design, planning, routing, and management of broadband communication networks, paying special attention to operation research techniques that can be applied in this area.

Jan Cheyns received the M.Sc. degree in electrical engineering (option micro- and optoelectronics) and the M.Sc. degree in physics from Ghent University, Belgium, in 1998 and 1999, respectively, where he currently is pursuing the Ph.D. degree.

In 1999, he joined the Optical Materials and Devices group of the Department of Information Technology. In 2000, he received a Fund for Scientific Research of Flanders scholarship and joined the Broadband Communications Network Group. His main interests are packet and burst switched optical networks.

Ilse Lievens received the M.Sc. degree in electrical engineering and the Ph.D. degree from Ghent University, Belgium, in 1994 and 2000, respectively.

Her Ph.D. thesis ("Use of Distributed Rerouting in Meshed ATM Networks") looked at the design of rerouting algorithms for survivability in meshed ATM networks, with a focus on distributed, autonomously working techniques. She is currently a Postdoctoral Assistant in the Broadband Communication Networks Group. Her research interests include reliability and survivability for ATM, IP, and optical networks. She is involved in the IST project LION and the Flemish interuniversity project "Design and Control of Broadband Networks for Multimedia Applications."

Mario Pickavet received the M.Sc. and Ph.D. degrees in electrical engineering from Ghent University, Belgium, in 1996 and 1999, respectively.

From 1996 to 1999, he was a Researcher in the Broadband Communications Networks Group, Department of Information Technology, Ghent University. Since 2000, he has been a Professor at Ghent University, where he is teaching telecommunication networks and algorithm design. His current research interests are related to broadband communication networks (IP, MPLS, WDM, SDH, ATM) and include design, long-term planning, and routing of core and access networks. He devotes special attention to operations research techniques that can be applied for routing and network design. In this context, he is currently involved in the European IST projects LION and DAVID.

Paul Lagasse received the M.Sc. degree in electronic engineering and the Ph.D. degree from Ghent University, Belgium, in 1969 and 1972, respectively.

After originally doing research in the areas of surface acoustic waves, high-frequency technology, and optoelectronics, he is now managing a research group of more than 100 academics working in the fields of telecommunications and information technology. He cofounded two spinoff companies in the areas of CAD software for high-frequency circuit design and of telecommunication software. Currently, he is also involved in assessing R&D policies and programs on a national level as a Member of the Board of the Flemish Institute for Science and Technology and recently on a European level as a Member of the external monitoring committee on the ACTS programme and of the "Advisory Group" of the IST programme. He has published more than 100 papers in national and international journals in the fields of microwaves, microwave acoustics, optoelectronics, III–V semiconductor technology, and broadband telecommunications.

Piet Demeester received the Ph.D. degree from the Department of Information Technology (INTEC), Ghent University, Belgium, in 1988.

He was with the Interuniversity Micro Electronics Center (IMEC) and became a Professor at Ghent University in 1993, where he is currently teaching telecommunication networks and is responsible for research on communication networks. He was Chairman of the First International Workshop on the design of Reliable Communication Networks in 1998 (DRCN98). He is a Member of the editorial board of the journals *Optical Networks Magazine* and *Photonic Network Communications*. His current interests are related to broadband communication networks (IP, ATM, SDH, WDM, access) and include network planning, network and service management, telecom software, internetworking, network protocols, etc.